Chapter 1

# Evolutionary Paths to Corrupt Societies of Artificial Agents

**Walid Nasrallah**
American University of Beirut, Lebanon
walid@alum.mit.edu

Virtual corrupt societies can be defined as groups of interacting computer-generated agents who predominantly choose behavior that gives short term personal gain at the expense of a higher aggregate cost to others. This paper focuses on corrupt societies that, unlike published models in which cooperation must evolve in order for the society to continue to survive, do not naturally die out as the corrupt class siphons off the resources. For example, a very computationally simple strategy of avoiding confrontation can allow a majority of "unethical" individuals to survive off the efforts of an "ethical" but productive minority. Analogies are drawn to actual human societies in which similar conditions gave rise to behavior traditionally defined as economic or political corruption.

## 1.1 Introduction

"The evolution of cooperation" or of "altruism" is an idea whose popularity has turned it into a bit of a cliché in the world of artificial life modeling. Briefly, a person, animal or computational agent needs to take resources away from the fulfilment of its own needs in order to contribute to the well being of others in its social unit. Doing so does not seem consistent with the imperative of natural selection, which rewards efficiency in using resources towards the fulfilment of a creature's own needs. But whenever the presence of others in the environment

is part of the basic needs of a creature, then purely self-serving behavior over a course of generations runs the risk of leaving the creature alone, and hence unable to continue to propagate. Therefore, a creature that somehow acquires the habit of helping others would over the long term enjoy an evolutionary advantage over a similar one that did not.

Since the same argument can be made for any member of a population, it stands to reason, and has been corroborated by simulations and biological observations, that cooperative behavior should be widespread in any population where social interaction is valuable. Of course, freeloaders and parasites can always come along, but they need to be eliminated, avoided or at lest kept in check in order to ward off extinction of the population.

This paper examines the incidence of what I call "corrupt societies", which are characterized by a prevalence of non-cooperative behavior in a social system that nonetheless continues to survive. These societies are studied from the abstract vantage point of iterated prisoners' dilemma interactions between computational agents. In particular, I look at five different ways in which a corrupt society can evolve: two of them from past literature, two observed by myself in the course of my research, and one that I expect to be discovered from future simulations. The latter would follow logically from the previous scenarios, but the exact details will, at least initially, depend on interactions best revealed by simulation. I conclude with some speculation about how the incidence of corruption (by the familiar economic/politcal definition) in contemporary human societies seems to correlate with factors analogous to the evolutionary paths in the simulation model.

## 1.2 Background

### 1.2.1 Iterated Prisoners' Dilemma (IPD)

Studied under "reciprocal altruism" in sociobiology and "eusociality" in entomology, the idea of "evolution of cooperation" found its purest mathematical expression in the "Iterated Prisoners Dilemma" problem. Characterized by the lack of a "Core", the original Prisoners' Dilemma of Game Theory provides each participant or player with two choices, to "cooperate" with the other participant or to "renege" on the implicit promise to cooperate. The payoffs are arranged to provide a temptation $t$ for an individual to renege when the other cooperates, a punishment $p$ when both renege, and a reward $r$ when both cooperate. The least favorable outcome is $s$ (for sucker) when a cooperator faces a reneging player; this makes it more advantageous for that sole cooperator to shift to the $p$.

A Nash equilibrium exists when both players defect, because a player that changes from defect to cooperate without any guarantee from the other player must lose, by going from $p$ to $s$. In addition to being an instinctively undesired state of affairs, the $p, p$ outcome is also not stable under the effect of a coalition between the two players, who would then move to the $r, r$ state if both cooperate. But then, the coalition can be undermined by either player yielding to

| | **Cooperate** | **Defect** |
|---|---|---|
| **Cooperate** | r<br>r | s<br>t |
| **Defect** | t<br>s | p<br>p |

**Figure 1.1**: Payoffs for classic Prisoners' Dilemma. $t > r > p > s$ and $2r > (t + s)$. (The column-chooser is underlined for clarity).

the temptation to obtain $t$ instead of $r$. Hence no outcome is stable under free coalition formation and destruction. In other words, the game has no *core*.

This changes when the game is played multiple times. Under multiple iterations, players can decide to punish a defector and reward a cooperator. In the long run, strategies based on cooperation can be shown to be more stable under an evolutionary regime where payoffs allow a player to continue to exist and to possibly give rise to copies of itself.

Within the general state where cooperation is favored, many different strategies can be dreamt up to maximize cumulative rewards. Different ways to remember past defections and to react to them can be pitted against different ways to sneak in an occasional defection among a string of cooperations. The defections give a temporary boost in reward, while the cooperations serve the dual purpose of keeping the opponent alive and of lulling against retaliation. The most famous outcome of allowing multiple strategies dreamt up by human researchers is reported by Axelrod, [2, 3]. It was found that the injunction to "Keep it simple, stupid" worked best: a strategy that simply cooperated the first time and then responded to every opponent by repeating the last action of the opponent was the winner against many other more sophisticated strategies. This strategy was descriptively and enduringly named"Tit-for-tat" (summarized *TfT*. Even strategies designed to fool "tit-for-tat", or to at least overcome the one defect of indefinitely repeated cycle of retaliation-atonement following a misunderstanding, were not able to survive contact with a multitude of other possible strategies as well as "tit-for-tat" itself [3].

### 1.2.2 Corruption in IPD and real-life analogues

"Tit-for-tat" is a strong guardian against corruption in a society. The strategy survives very well and continues to immediately detect and respond to the slightest attempt to swindle others by pretending to cooperate for a while. This is done without any large investment in memory, since both free-loaders, who try and sneak in occasional defections, and other retaliators, do not attain greater success even when they have a memory of many past interactions. The simplicity of the winner of Axelrod's competitions inspired [8] to scour the complete strategy space consisting of one-encounter-deep memory of both the agent's own move and that of the opponent. It was found that another simple strategy can be even more successful. "Pavlov" was the name given to the strategy after

it was found to be more evolutionary persistent than initially expected by its initiators, who had originally named it "Simpleton". "Pavlov" (summarized as *Pav* does not cooperate when it sees a cooperator, but instead repeats its own last action. Similarly, when it sees a defection, it does not defect, but rather switches its past action to the opposite move. Under random evolution, the ability of "Pavlov" to exploit a partner that always cooperates is of more value than "tit-for-tat's" ability to persistently retaliate against a partner that always defects.

Applied to human interactions, *TfT* reminds one of the reluctant gunfighter of movie westerns, or of the citizen-soldier creed of America's George Washington, Rome's Lucius Quinctius Cincinnatus, and Homer's Ulysses. "Pavlov" is an amoral opportunist, more like Homer's Agamemnon or any of a number of more recent politicians. In any case, the presence of either or both *TfT* and "Pavlov" provides the immediate means to reduce the initial success of habitual defectors. In the long term, the presence of someone who can retaliate and survive, is necessary to prevent the extinction of the whole society, which would otherwise be the less direct consequence of the initial success of defectors within a society.

## 1.3 Previously Studied Paths

What, then, allows a society to be overcome by people or agents that defect more than they cooperate when different types of retaliators are always around? The answers are still a topic of debate, since simulations that produce an example of a corrupt society can be shown to be special cases of a general situation where defection remains an evolutionarily losing strategy.

### 1.3.1 Fast Predators

Since agents only learn from their past experience, a habitual defector can thrive if it can seek new victims more quickly than it can be tracked down by those who would retaliate against it [4, 5]. An equilibrium ratio of fast-moving defector to other players including *TfT* can be calculated from the speed of movement, the density of agents, and the size of the "patch" or immediate neighborhood within which agents can see and interact with each other [4]. This type of model raised worries that cooperation occurs more in the real world of animal, plant and human interactions than in the abstract simulation. This was explained by factors from outside the original model, such as fast spread of information between agents about who is a defector and a pre-existing assumption that a stranger is a defector until otherwise proven [5]. Augmenting the model in that way resolves the immediate issue of disparity between the model and observations, but it raises the possibility of second-order strategies such as false information, camouflage, and other strategums observed in the animal kingdom. What we have is one path, mostly of metaphorical value, to a society that can indefinitely maintain a large number of defectors.

In a human society, his theoretical path to corruption can be compared to mounted nomadic raiders ravaging medieval peasants' fields, or militiamen on pick-up trucks assailing contemporary villagers in a failed nation.

### 1.3.2 Watered-down punishment

It is also possible to play with the payoff matrix of the interactions to lower the cumulative cost to society of defection — i.e. increase $s + t$ visà)-vis $2r$ [7]. When the value of $t$ in the payoff matrix is between 1.85 and 2, and $s$ and $p$ are 0, and $r$ is to 1, then a society can evolve with a majority of defector. In addition, deterministic rules for changing strategy from "defect" to "cooperate" and vice versa gives rise to a chaotic or random shifting pattern in which the ratios are roughly maintained. Although this simulation was done with immobile agents that spread their strategy to immediate neighbors like lichens or trees, the conclusion that parasitic behavior can flourish when its ill effects are limited is a useful analogy in other fields.

One analogy in a human society might be with low-intensity interactions, such as queue-jumping or tax evasion. The range of real behaviors that can be modeled with such a restricted subset of possible interaction payoffs is small, but the metaphor has its place in characterizations of human social behavior.

## 1.4 Corruption among equally mobile agents

A series of simulations [1] was conducted to test what happens when mobility is no longer the exclusive to defectors. It was found that a cooperator that can avoid defectors is much more successful than across a wide range of conditions, including density, payoff matrix differences and mobility. In a human society, this corresponds to the familiar trait of social opprobrium. A school bully does not get invited to birthday parties; a politician who successfully evades the law loses the next election. The predominance of cooperation is reasserted in a metaphor for human behavior in which all have equal mobility and some level of heterogeneity exists among the population.

### 1.4.1 High-stakes anonymous interactions

The same strategy that helps a society avoid becoming corrupt when mobility is introduced can, in some situations, become itself a source of a new form of corruption [6]. Although [1] showed that cooperation predominates, there are distinct areas of the model's parameter space where the mobile defectors come to outnumber the other mobile strategies. Typically, in the few instances when this happens, the *TfT* and *Pav* agents are driven to extinction, and a mobile cooperator that avoids defectors without retaliating becomes the source of productivity in the population while remaining numerically in the minority.

The type of corrupt society arises when density is high, mobility is high, and agents interact repeated times before disengaging and looking for a new

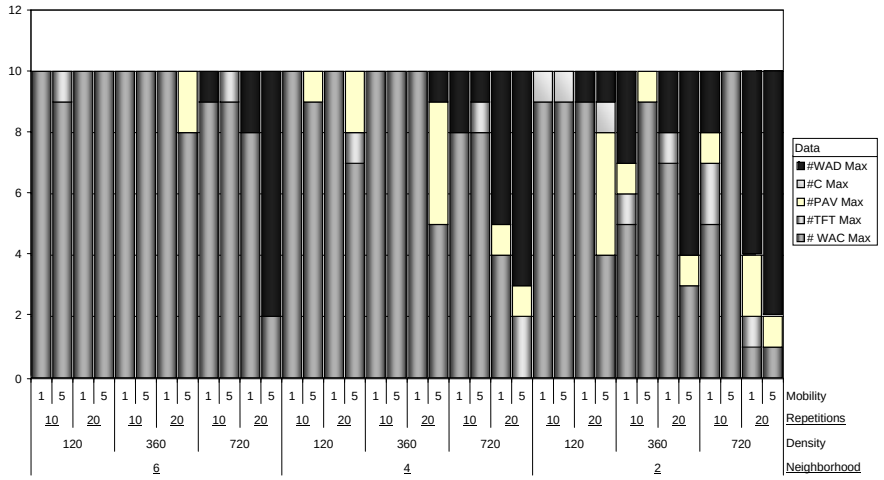


**Figure 1.2**: Dominant strategies with anonymity [6]

partner. This means that meeting a defector can lead to repeated exposure to defection, which has the effect of weeding out the retaliators in the first few hundred simulation cycles. In addition, agents do not recall the identity of their last inte4raction partner, so a re-encounter with the same defector is met with initial cooperation even by retaliators. This may seem far-fetched metaphor for human interactions. However, the evolution of the internet has given rise to a level of anonymity in many high-stakes interactions that this particular abstract path to corruption can potentially teach us something about what to look for and what to avoid in planning human some certain social actions.

The "Nigerian 419 scams" [10] are one example that springs to mind of a human social phenomenon becomes prevalent due to this underlying dynamic.

### 1.4.2 Dense, fragmented societies

In two figures Fig. 1.2 and Fig. 1.3, the dark bars at the top show the proportion of simulations where a defecting avoider has the highest numbers after 1000 runs, i.e. where the resulting society can be said to be corrupt. The most prominent observation is that the ability to detect when a new encounter is with an agent previously seen leaves a possibility of corruption only when small neighborhood size is combined with high density. This combination has the same effect as §1.3.1 above. A predator always find more victims in the high density population, but victims cannot hide as easily because of the low neighborhood size. (Interactions are only possible within the patch or neighborhood unless the agent moves away.) The interesting thing is that the predators do not need to be more mobile than other agents in order for this effect to be seen.

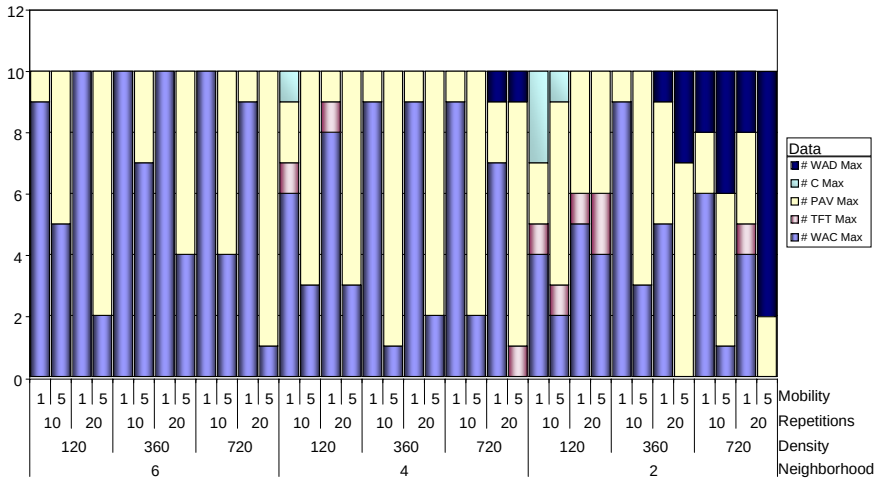


**Figure 1.3**: Dominant strategies with 1-partner-deep memory [6]

## 1.5 Kinship Bias: A future path for research

§1.4.2 described how small neighborhoods and high density, if enforced as physical constraints, lead to a higher preponderance of corrupt societies. One human analogue to this abstract path to corruption is ethnic-based ghetto-like communities in an urban setting. Today, most ethnic neighborhoods in a multi-ethnic city are self-selected, not enforced. This does not make much of a difference when mapping something so abstract as an IPD simulation with memoryless mobile agents to a to human social experience: one more feature not in the model. But one can wonder how the incidence of corruption changes when the agents reduce their pool of immediate interaction candidates in a different way more analogous to un-integrated ethnic sub-populations.

From an evolutionary standpoint, it is widely accepted that costly altruistic behavior is more advantageous when directed at kin. IN the context of IPD, predicting when a newly encountered agent will be a defector can boost survival, and one way to do this is to look at how the prospective partner shares ancestors with other previously encountered agents or with the agent doing the evaluation itself. It has been suggested [9] that human societies whose members value kinship are more likely to become weaker than societies whose members treat each other equally. The increased corruption due to the small-neighborhood effect might be one way to explain this idea in abstract dynamic terms. Future research where the kinship values are directly coded into agents' strategies will be needed to show whether other, possibly advantageous, effects of kinship valuation can increase or decrease the incidence of corruption and hence the susceptibility of the society to invasion by a less corrupt group.

# Bibliography

[1] Aktipis, C. Athena, "Know when to walk away: contingent movement and the evolution of cooperation", *Journal of Theoretical Biology* **231**, 2 (2004), 249–260.

[2] Axelrod, Robert, "More effective choice in prisoner's dilemma", *The Journal of Conflict Resolution* **24**, 3 (September 1980), 379–403.

[3] Axelrod, Robert, *The Evolution of Strategies in the Iterated Prisoner's Dilemma*, Morgan Kaufman, Los Altos, CA (1987), ch. 2.

[4] Dugatkin, Lee Alan, and David Sloan Wilson, "Rover: a strategy for exploiting cooperators in a patchy environment", *American Naturalist* **138**, 3 (1991), 687–701.

[5] Enquist, Magnus, and Olof Leimar, "The evolution of cooperation in mobile organisms", *Animal Behaviour* **45**, 4 (1993), 747–757.

[6] Nasrallah, Walid Fawzi, and Youssef George Saad, "The role of partnering velocity and spatial mobility in the evolution of cooperative and parasitic behavior", *Adaptive Behavior Journal* (2006), Under review.

[7] Nowak, Martin, and R. M. May, "Evolutionary games and spatial chaos", *Nature* **259**, 6398 (1992), 826–829.

[8] Nowak, Martin, and Karl Sigmund, "A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game", *Nature* **364**, 6432 (1993), 56–58.

[9] Peters, Ralph, "Spotting the losers: Seven signs of non-competitive states", *Parameters* **28**, 1 (1998), 36–47.

[10] Zuckoff, Mitchell, "The perfect mark : How a massachusetts psychotherapist fell for a nigerian e-mail scam", *The New Yorker Magazine* (May 15 2006).